Unsupervised Segmentation & Correspondence

Rehan Ahmad

December 2024

1 Abstract

Human vision is *compositional*. From just a few examples, we are able to naturally break down objects into hierarchical components based on their structure. For example when viewing a car, we intuitively perceive it as a whole but can also identify its components such as its wheels and doors. The elegant feature of these compositional representation is that we develop these ***categories*** *without explicit guidance*. We rely instead on repeated exposure and inherent cognitive mechanisms to recognize objects within a category.

Recent advances Text-to-Image diffusion models show remarkable capacity in learning an **unsupervised** representation of data. The process of gradually denoising the images by predicting the score function $s_\theta(x_t|c)$ across $t = 1, ...T$ timesteps provides a strong estimate over the natural distribution $p(x)$ of images conditioned on c where c represents textual prompt. Although powerful, this representation lacks the compositionality associated with human vision. This is a setback because compositionality is an appealing feature encountered in many computer vision tasks. For example, in **keypoint detection**, compositionality is essential for understanding articulated objects such as identifying the pose of a person in a 2D image. Similarly, in **object occlusion**, compositionality allows a system to infer the presence of partially hidden objects by recognizing visible components and reconstructing the missing parts based on learned structural relationships.

Given the invaluable role of compositionality in the domain of computer vision , we aim to extract the relevant signals from the over-parametrized diffusion pipeline to learn this characteristic in an **unsupervised fashion**. To carry this out, our hypothesis is that a model capable of performing **unsupervised segmentation** and **correspondence** inherently demonstrates compositionality as it learns to decompose objects into meaningful parts and establish relationships between them despite variations in pose, lighting, or occlusion. Furthermore, we aim to test this compositionality by placing the onus on the diffusion model to "imagine" the entire objects from its part - thus, demonstrating the utility of such a representation.

Given a Dataset $D = \{(x_1, x_2, , ..., x_n\}$ representing objects of similar categories such as birds, we leverage the capabilities of self-attention and cross-attention maps to extract meaningful signals related to these objects. In particular, the ability of the self-attention map $A_{self}[i, j]$ for pixel (i,j) to capture other pixels belonging to the same group naturally makes them suitable for the task of grouping together objects and segmenting parts $p_1, p_2, ...p_n \in x_i$.

Seeing a diverse range of objects within a category (e.g., birds of different sizes, colors, and shapes) helps humans identify the critical features that define the category. For instance, seeing birds with and without colorful plumage emphasizes that features like beaks and wings are more fundamental to the "bird" category than feather color. This important ability to converge to representations

that consistently explain the objects across $x_1, x_2, ..x_n$ motivates us to seek correspondences between segmented regions. To carry this out, we use textual inversion to train a collection of embeddings in an unsupervised fashion $c_i, ...c_n$ to converge to representations that consistently explain the Dataset.

While previous works by *E Hedlin et al* carried out textual inversion to establish local correspondences, we show that the loss objective introduced in their formalism is ineffective in approximating the underlying category of interest which is critical for generative purposes. Thus, we resort to the stronger supervisory signal of noise-estimation to get a useful approximation of the underlying CLIP-embedding explaining the categories $p_1, ...p_n$.

Finally, we use these collection of embeddings to solve the problem of occlusion. In particular, given an occluded object, we aim to pinpoint which categories are missing and then "imagine" the occluded object using the pre-trained embeddings. We hope to update the full abstract by the end of January, and will release pre-results by the end of February.